A Signaling Primer

1 Introduction

The intention of this short treatise is to discuss some alternative connection methodologies for propagating signals between modules (typically integrated circuits) on boards or within systems. The distances traveled may be short, for example, a few centimeters on boards but could range up to several meters in systems. Electrical signals are considered here and not photonic (light) signals.

To keep this discussion simple, a signal is defined as having a source or *transmitter* and a destination, or *receiver* and travels along a path, typically a wire. (More complex signaling may entail multiple sources and destinations and is not discussed here.) A signal propagates information from the transmitter to the receiver. This information is encoded as a time-varying voltage or current. The time-varying property is often called 'alternating current' or 'AC'.

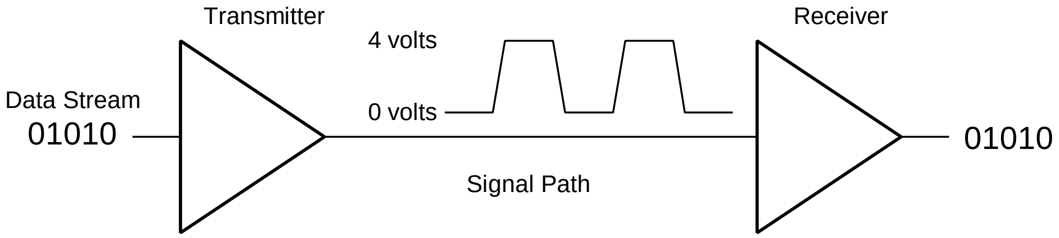

Figure 1: A transmitter sending a time-varying waveform to a receiver.

The speed of signal propagation is (nearly) the speed of light, which is quite fast. However as data rates become high, propagation speed can interact with the distance the signal must travel and this becomes a factor in designing the path. The end result is that a signal cannot be presumed to travel instantaneously, but rather, the transmitter is somewhat 'ahead' of the receiver. It is even possible for a signal to be 'stored' (briefly) on a path, as it travels from transmitter to receiver.

Signals can be damaged such that the encoded information is degraded or lost. This is typically an 'additive' process, where an unwanted (noise) signal is added into the information signal. It is difficult to keep noise sources out of signal path so a good deal of design work goes into controlling the amount of noise that gets in. Typically, if the amplitude of the noise is small compared to the amplitude of the signal, the encoded information can be extracted by the receiver without loss.

Digital systems encode data as streams of 0s and 1s and this data is represented (typically) by a time-varying signal where a 0 is a lower voltage (say, 0 volts) and a 1 is a higher voltage (say, 4 volts). A transmitter will send a time-varying voltage out on a wire. This signal will propagate and some time later, the receiver will sense it and recover the signal. If a small amount of random noise voltage (say +/- 0.5 volts) is added to this signal during its propagation, the receiver will sense this noise as well. However, if the receiver is designed to interpret any voltage above a certain *threshold* (say, 2 volts) as a 1 and any voltage below the threshold as a 0, then this noise voltage will not affect the transmission of digital data. Note it is extremely important that both the transmitter and receiver are in agreement on what the data and threshold voltages are. If a transmitter is sending a signal with a swing of 0 to 4 volts, but the receiver has a threshold at (say) 0.5 volts, then random noise on the signal may be interpreted as part of the information.

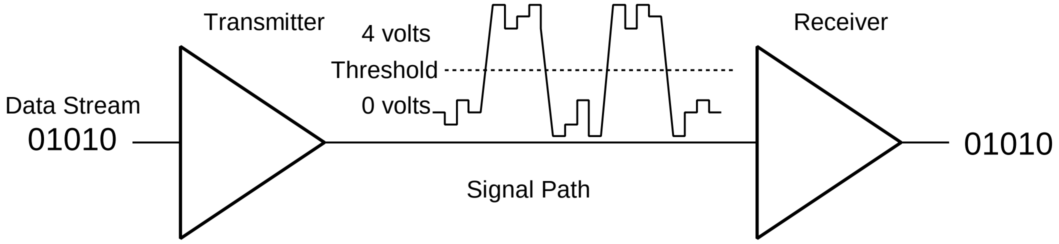

Figure 2: Transmitted signal with a small amount of random noise added in.

A signal contains *power* which is to say it takes energy to transmit information. This power is a necessary feature, but can lead to power dissipation problems, namely *heat*, that affect our design considerations. Typically a designer wants to keep the power needed by a system within design limits. One way to do this is to limit the power in signals.

Power can be limited by slowing down the rate of information transfer, or by limiting the *voltage swing* of the signal. (The power consuming properties of the signal path itself can also be controlled with some effort.) It turns out that with all other things remaining the same, the power in a signal is *linearly* related to the rate of data transmission and to the *square* of the voltage swing. That is to say, increasing the data rate by two increases the power required by two, but doubling the voltage swing quadruples the power required. Since devices and systems always seem to go faster with each new generation, we see data rates going up, increasing power demands to levels that are troublesome to designers. Thus they react by lowering the voltage swings of the signals. However this is fraught with risk, since as voltage swings become smaller, the noise immunity offered by signal thresholding is lowered as well. This can lead to data errors and other reliability problems.

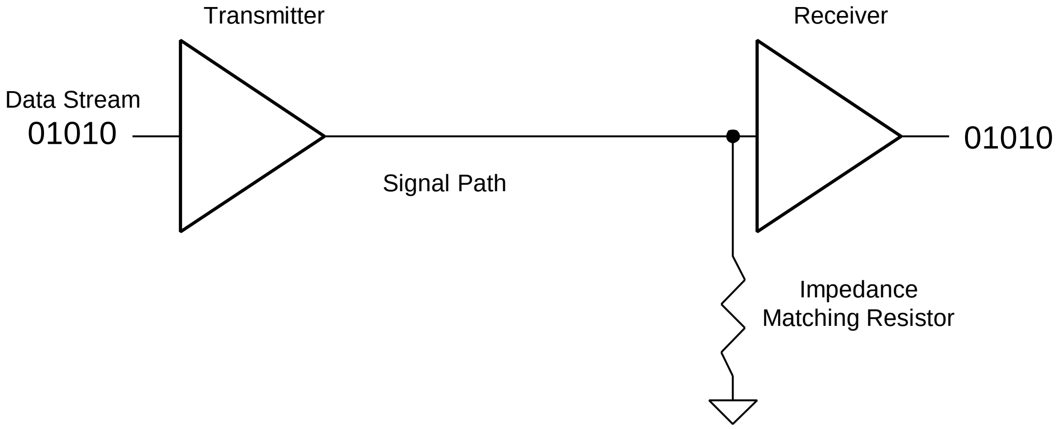

Figure 3: A receiver with an impedance matching resistor.

A signal path also has something called *path impedance* which is a measure of its reaction to a swiftly changing AC signal. Path impedance is a characteristic that is only important when data rates become high which is certainly the case now. Just a few paragraphs back it was mentioned that a signal could be ‘stored’ momentarily on a wire. Even more surprising is that this *moving wave* can literally *reflect* off the end of the wire (by the receiver) and bounce backward in the opposite direction. This reflected wave (only some of the energy reflects, so it is a reduced version of the original) now adds into the subsequent data waveforms generated by the transmitter, which is just like adding in noise. This phenomenon can be canceled out by something called *impedance matching* which is performed by adding a resistor (or network of resistors) to the end of the line. This resistor value is selected to equal the path impedance. It absorbs the incoming signal rather than allowing it to reflect. Notice the other end of the resistor is connected to a stable voltage reference, in this case ground.

2 AC Coupling

So far, DC coupling, a directly wired connection from transmitter to receiver has been considered. However, AC coupling is becoming important because it helps solve some problems. It is usually performed by adding a capacitor to the path. This capacitor is sometimes called a *DC blocking capacitor* because it prevents the flow of DC current, in this case, from the transmitter to the receiver and matching resistor.

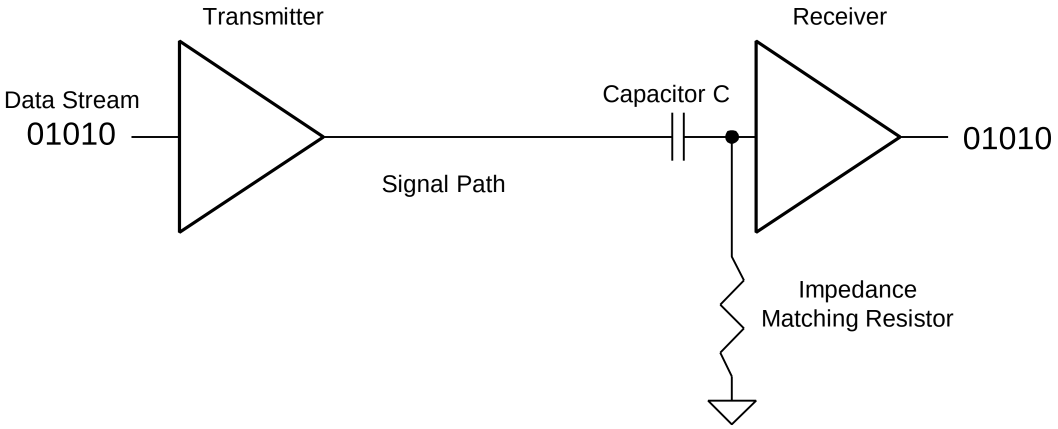

Figure 4: An AC coupled signal path.

2.1 AC Signals

The combination of a series capacitor and the impedance matching resistor forms what is known as a *high-pass filter* which is a circuit that passes higher frequency AC signals, but reduces those of lower frequency. In fact, DC has a ‘frequency’ of 0, so DC is completely blocked. Thus if a transmitter is sending rapidly varying waveforms (high frequency) these will pass through the filter almost unimpeded. If the transmitter was also sending an added DC value, this is blocked.

In previous examples, the transmitter was sending either 0 or 4 volts. The *average* voltage over time is somewhere above 0 volts depending on the amount of 1s and 0s in the data stream, so there is an average DC voltage over time.

Figure 5 shows a transmitted waveform and what is seen on the other side of the filter. Notice that the rising edge of the waveform passes the filter, but the received voltage then begins to decay to zero. This is because the capacitor has charged up on the initial edge. Then the first falling edge occurs, which causes a negative-going downward pulse that also decays. The frequency of this waveform is low enough that complete charging of the capacitor occurs between edges.

Figure 6 shows the same high-pass filter with a much higher frequency waveform applied. The same process of transmitting an edge and subsequent decay is still visible, but since the edges are much closer together, much less decay is evident. Note that the received voltage swings from (approximately) 2 volts to –2 volts where the input waveform moved from 4 volts to 0 volts. The swing (approximately 4 volts) is preserved, but the average DC value is zero. So, two important effects are evident: first, the higher the frequency of the waveform, the more preserved its appearance after the filter. Second, the DC component of the input waveform has been removed.

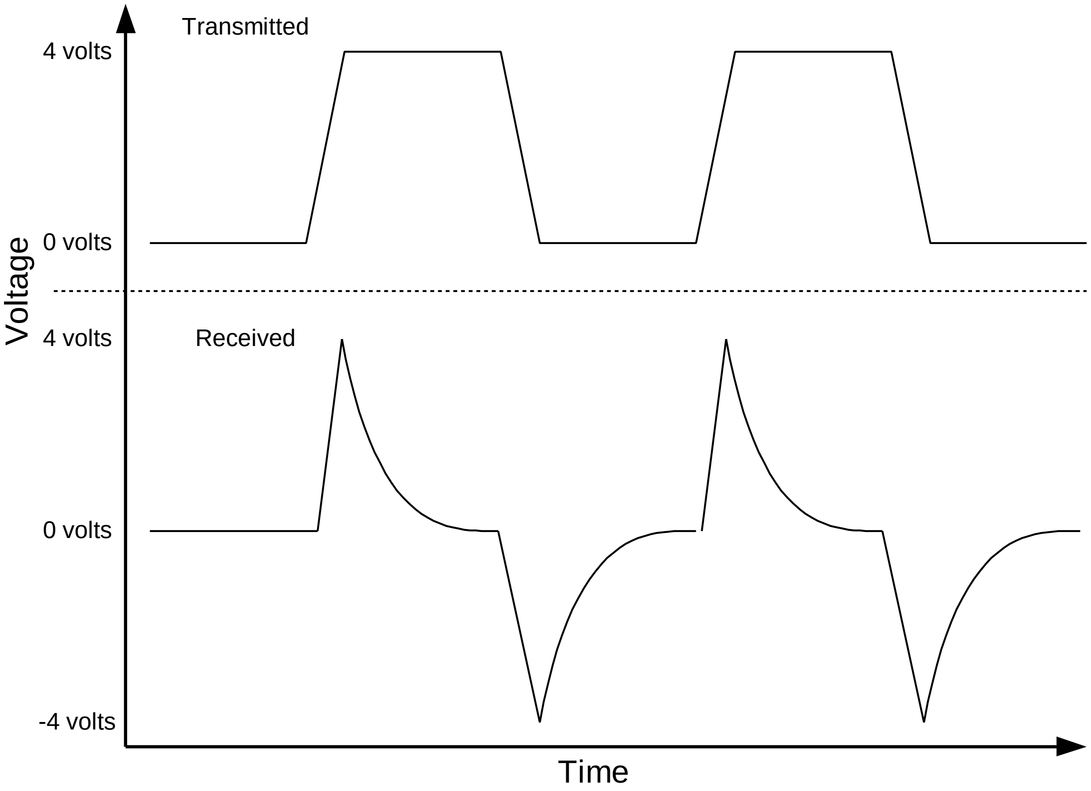

Figure 5: Two voltage waveforms, on either side of the blocking capacitor.

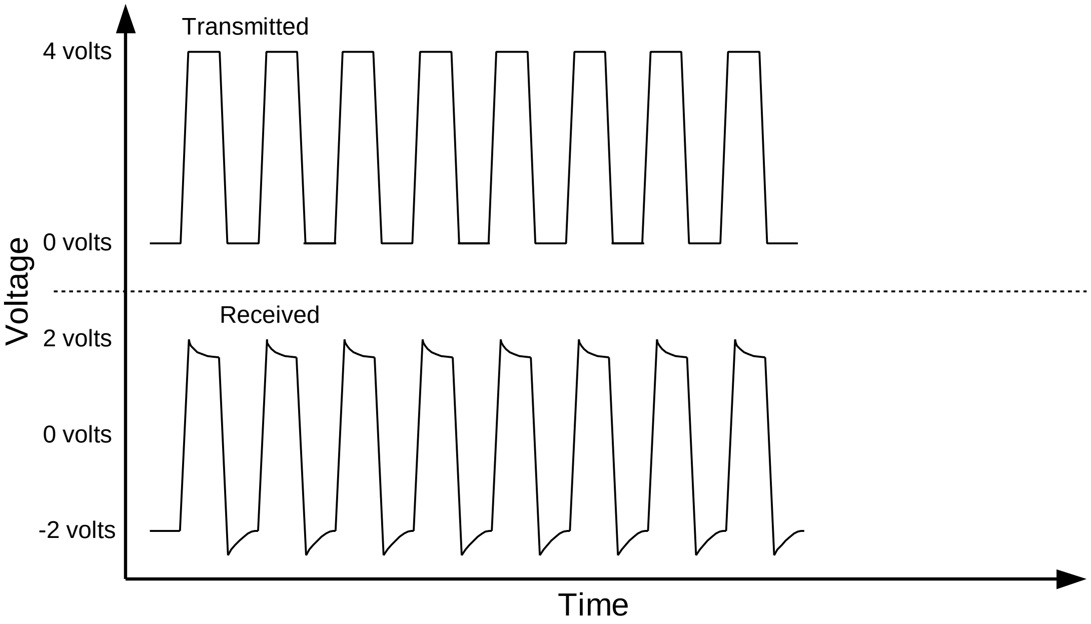

Figure 6: Higher frequency waveform, before and after the capacitor.

2.2 Transient and Steady-State Response

The behavior of the high-pass filter has one more detail to consider. The previous figures showed how the filter behaves after a large number of cycles of alternating data bits are transmitted. This is the *steady-state response*, and the steady-state waveform was centered about 0 volts. However, if the waveform was recently ‘turned on’ as shown in Figure 7, we see there is a time where the voltage is not centered at 0, but effectively centered at 2 volts. This is the *transient response*.

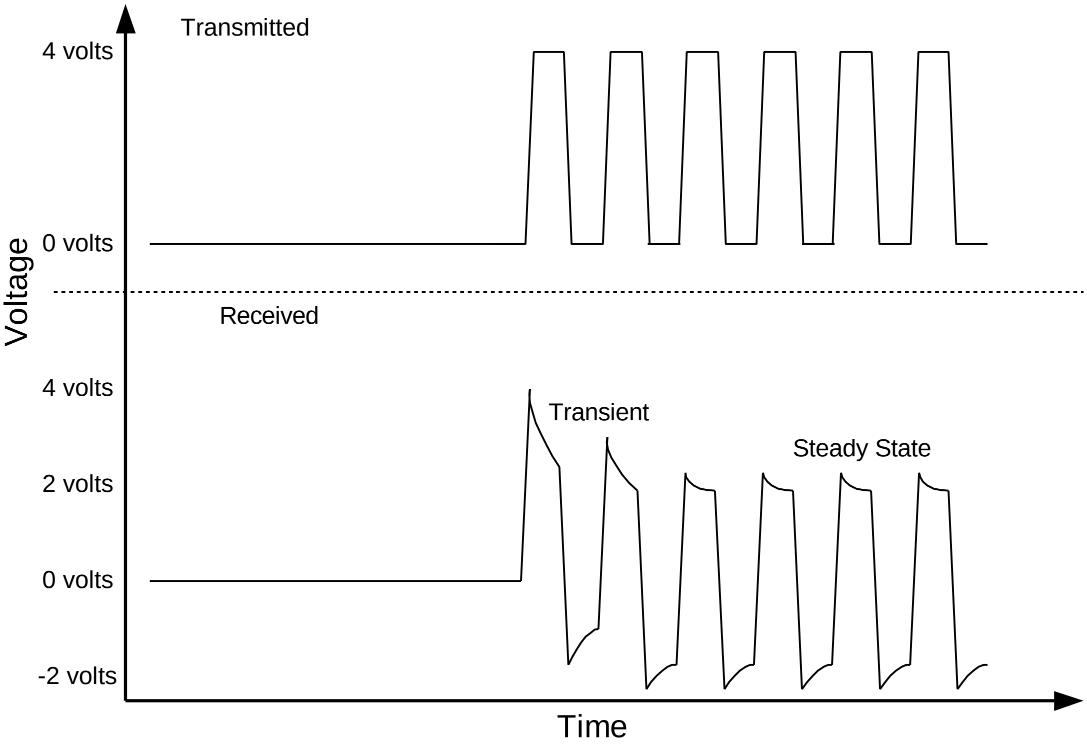

Figure 7: Transient and steady-state response of the signal path.

The transient response is now an important consideration. If the receiver is designed to expect a valid signal that is centered near 0 volts, then it may not operate correctly when it sees a signal that contains a transient response not centered there. In the example above, the reason that the transmitted data was initially 0 is that the data stream being transmitted contained a long string of 0 data. An imbalance of 0 (or 1) data could cause a transient response. Designers of high speed data signals try to avoid this problem by sending extra bits in the stream that compensate for imbalances in the data. These bits are located in known positions and are later ignored. They are there simply to *condition the signal* appropriately for the electronic characteristics of the signal channel. So these extra bits are added in with foreknowledge of the data stream, to provide better balance and reduce transient responses.

2.3 DC Offsets

When DC blocking capacitors are used in signal paths, this can allow normally incompatible drivers and receivers to be used together. For example, if a driver produces enough voltage swing to be recognized by a receiver, but the threshold voltage is incompatible, AC coupling and the proper termination voltage at the receiver can reestablish compatibility. This allows designer more freedom in choosing devices that must work together.

3 Differential Signaling

All of the signaling seen so far have been *single-ended*, meaning there was a single pathway sending a stream of data bits. Single-ended signaling depends on signals having certain parameters, such as a threshold and voltage swing, that are constant. For example, the previous examples used a transmitter with a 2 volt threshold and a 4 volt swing (0 to 4 volts). The receiver had to match up to these parameters. The receiver often can accept a much smaller swing (say 2 volts), but the threshold is quite important in how it interprets incoming voltages. The threshold is built into the design of both the transmitter and receiver and is usually fixed. As already noted, noise sources added into signals can cause data to trip over (or under) the threshold producing errors. A signaling scheme that depends on fixed thresholds will have this weakness.

One approach to reducing sensitivity to noise is to have *two thresholds*. One, the *low threshold*, governs when a signal is 0, and the other, the *high threshold*, governs when a signal is 1. The two thresholds are separated by a gap sometimes called the *forbidden zone*, where a signal may be interpreted by a receiver as either a 1 or 0 unpredictably. Thus a signal that is 0 is *below low* and a signal that is 1 is *above high* in threshold terms. A signal in transition must pass through the forbidden zone. This is done as quickly as possible to minimize the time a receiver is uncertain about the data on a signal. However, there is some time uncertainty about when the receiver will perceive the transition, as depicted in Figure 8.

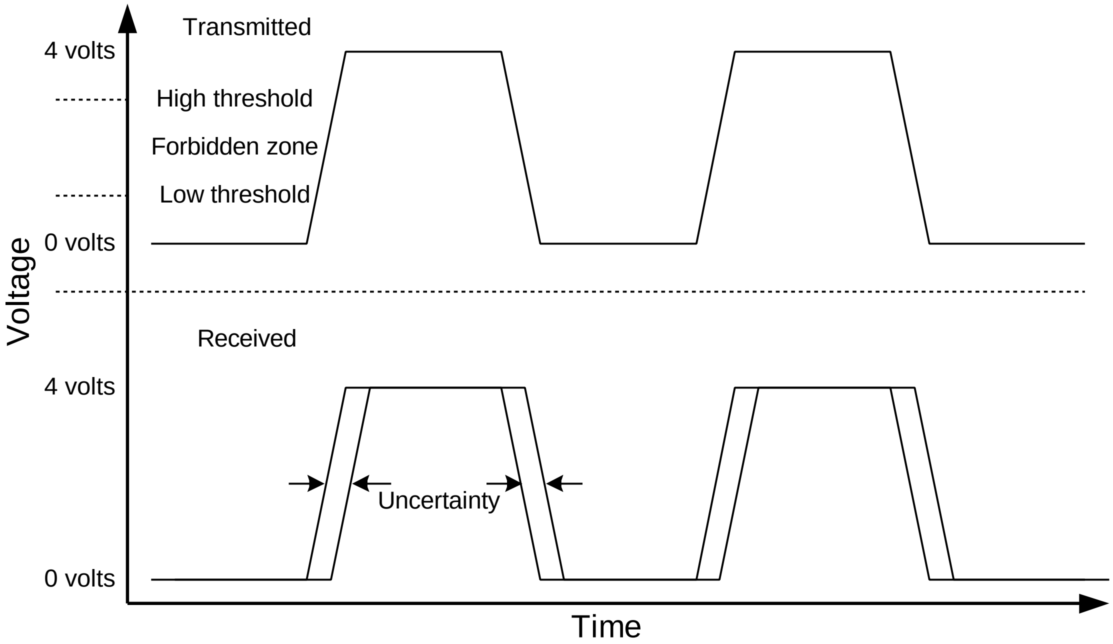

Figure 8: Dual-thresholds improve noise immunity, but add uncertainty to when a transition is received.

There is a second signaling technology in use called *differential* signaling. This scheme uses *two* wires to transmit the same data stream. One wire is the *positive leg* of the path and the other is the *negative leg*. There are two big differences. First, the two wires carry inverted copies of the voltage waveforms, meaning when one wire has a rising edge on it, the second has a falling edge. The second is that the receiver *compares* the two signals, producing a 1 whenever the signal on the positive leg is greater (more positive) than that on the negative leg. It produces a 0 when the opposite is true. Another way of saying this (which justifies the word ‘differential’) is that the two signals are subtracted (positive – negative) and a positive result is a 1 and a negative result is a 0. An example is shown in Figure 9.

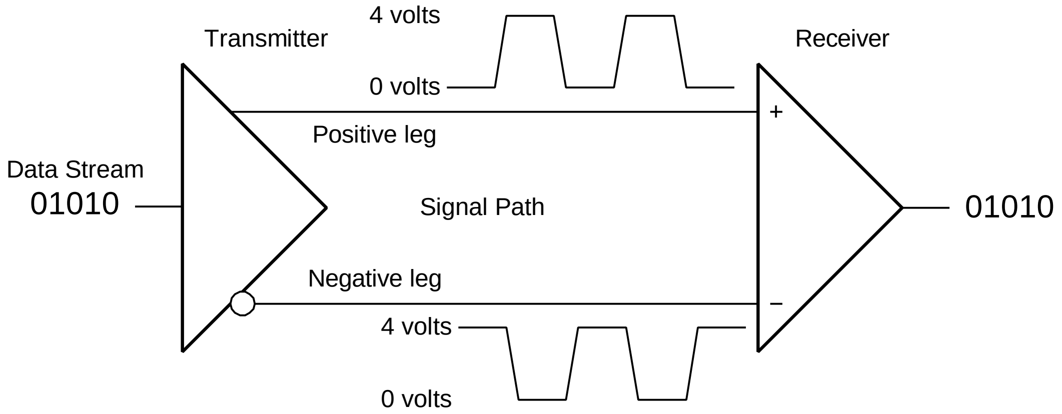

Figure 9: A differential transmitter/receiver pair.

3.1 High Speed Signals

As signal speeds continue to go higher, this means there is less time between edges in a signal. However the placement of edges is potentially uncertain as was seen in Figure 8. As you may guess, these two facts are in collision. We want to pack more transitions into a small time, but the uncertainty that is there may cause errors in the receiver.

Differential signaling helps in two ways. First, since the two signals are *mutual references*, there isn't a fixed threshold and the forbidden zone is greatly reduced. This means we are much more sure of when a transition will be perceived. Only when the two signals are nearly equal (in the middle of a transition) is there any uncertainty. Second, we can reduce the actual voltage swing. The effect of this is a transition has less 'distance' to travel in the same period of time. Thus we can achieve higher data rates with the same edge speed. Another way of viewing this is that the signal can change (called the *slew rate*) only so fast. Therefore, if the amount of the voltage swing is reduced (the 'distance' it must travel) then a signal can transition from 0 to 1 or vice versa much more quickly. This reduces power and increases the number of edges we can pack into a time period – and increases the data rate.

3.2 Noise Immunity

Differential signals have much better noise immunity. This can be seen by thinking mathematically about what is happening. Say the positive leg has a value of P and the negative leg has a value of N. Then the sign of the subtracted signals P – N is the result of the signal. Now let's add a noise error E. This gives us P+E *and* N+E, because most noise sources cause *common mode* noise, or noise that adds (nearly) equally into both legs of the signal. Now when the receiver subtracts the noisy signals, it gets (P+E) – (N+E) = P – N, as before. What is very interesting about this is that E may be large (even larger that P or N) and this cancellation still occurs!

4 AC Coupled Differential Signaling

Using differential signaling does have costs. Of course, two wires (and 2 more IC pins) are now needed to transmit a data stream where a single wire sufficed before. It is also a bit more complicated to design differential drivers and receivers. As design requirements for data streams with gigabit capacities become commonplace, we find this also complicates the design. As a result, to go from (say) 2.5 gigabits to 3.125 gigabits or higher, we need to adopt a new integrated circuit process technology. This in turn means different power supply voltages, different signal swings, different signal rise times, etc. As each IC vendor confronts these challenges, they may come up with somewhat different solutions. This can lead to ‘single-source’ problems for board and system designers; they find little choice available for compatible IC drivers and receivers. AC coupling, as shown in Figure 10, can help with these problems.

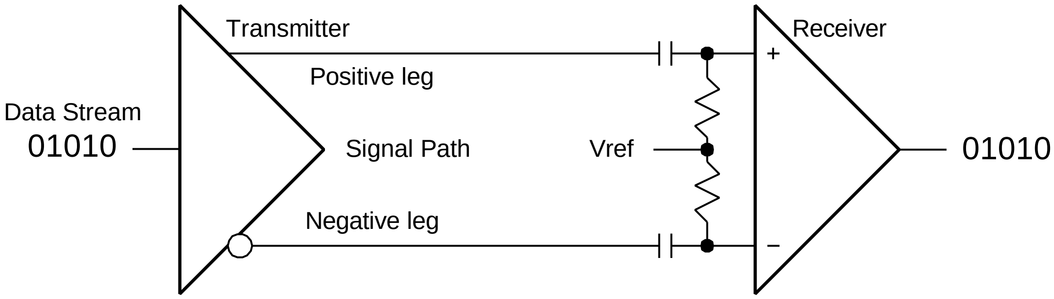

Figure 10: AC coupled differential signaling.

4.1 Increasing Compatibility

A transmitter from vendor A and a receiver from vendor B may be incompatible when DC coupled, but can be made to work when AC coupled. For example, in Figure 10 the *only* compatibility between transmitter and receiver may be the voltage swing that the transmitter creates on the path legs. The DC blocking capacitor, the choice of Vref and the termination scheme may then allow this configuration to work.

4.2 System Requirements

The configuration of Figure 10 is also ‘hyperactive’ in that it needs to see a continuous stream of data, that is, it can’t be ‘turned off’. Consider for a moment what happens if the driver stops transmitting data. Very soon after that, the capacitors completely discharge and then the value of Vref appears on both inputs of the differential receiver. It then subtracts (Vref – Vref) and produces a digital response. However the value of 0 (the difference) is in the uncertainty region, so the receiver will (probably) do indeterminate things as a result. Thus, to avoid this, the receiver always need some sort of signal, even when the system is ‘idle’. As before (see 2.2, Transient and Steady-State Response) the transient response may again cause requirements for ‘balancing’ the data streams seen by the differential receiver.

The gain is that AC coupled differential signal paths can operate at lower signal voltages (greatly reducing power consumption and heat), higher operating frequencies, with higher noise immunity. This satisfies several needs for high-speed system design.